

Form PTO-1390  
P21834-P01U.S. DEPARTMENT OF COMMERCE  
PATENT AND TRADEMARK OFFICE

ATTORNEY'S DOCKET NUMBER

P21834

TRANSMITTAL LETTER TO THE UNITED STATES  
DESIGNATED/ELECTED OFFICE (DO/EO/US)  
CONCERNING A FILING UNDER 35 U.S.C. 371

U.S. APPLICATION NO. (If known, see 37 CFR 1.5)

10/049627

INTERNATIONAL APPLICATION NO

INTERNATIONAL FILING DATE

PRIORITY DATE CLAIMED

PCT/SG99/00089

25 August 1999

TITLE OF INVENTION

DOCUMENT CLASSIFICATION APPARATUS

APPLICANT(S) FOR DO/EO/US



AR Hwee TAN and Fon Lin LAI

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information.

1. ☒ This is a FIRST submission of items concerning a filing under 35 U.S.C. 371.
2. ☐ This is a SECOND or SUBSEQUENT submission of items concerning a filing under 35 U.S.C. 371.
3. ☒ This is an express request to promptly begin national examination procedures (35 U.S.C. 371(f)).
4. ☒ The US has been elected by the expiration of 19 months from the priority date (PCT Article 31).
5. ☒ A copy of the International Application as filed (35 U.S.C. 371(c)(2))
  - a. ☒ is attached hereto (required only if not communicated by the International Bureau).
  - b. ☒ has been communicated by the International Bureau.
  - c. ☐ is not required, as the application was filed in the United States Receiving Office (RO/US).
6. ☐ An English language translation of the International Application as filed (35 U.S.C. 371 (c)(2)).
7. ☐ Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. 371(c)(3))
  - a. ☐ are attached hereto (required only if not communicated by the International Bureau).
  - b. ☐ have been communicated by the International Bureau.
  - c. ☐ have not been made; however, the time limit for making such amendments has NOT expired
  - d. ☐ have not been made and will not be made.
8. ☐ An English language translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371(c)(3)).
9. ☒ An oath or declaration of the inventor(s) (35 U.S.C. 371(c)(4)).
10. ☐ An English language translation of the annexes to the International Preliminary Examination Report under PCT Article 36 (U.S.C. 371(c)(5)).

## Items 11 to 16 below concern other document(s) or information included:

11. Assignee: KENT RIDGE DIGITAL LABS of SINGAPORE \_\_\_\_\_
12. ☐ An Information Disclosure Statement under 37 CFR 1.97 and 1.98.
13. ☐ An assignment document for recording. A separate cover sheet in compliance with 37 CFR 3.28 and 3.31 is included.
14. ☒ A FIRST preliminary amendment  
☐ A SECOND or SUBSEQUENT preliminary amendment.
15. ☐ A substitute specification.
16. ☐ A change of power of attorney and/or address letter.
17. ☒ Figure of Drawing to be published 3
18. ☒ Other items or information:  
Cover Sheet and International Application as published.  
PCT/IB/308.  
PCT/ISA/409  
PCT/ISA/210  
Cover Letter under 35 U.S.C. 371 and 1.495.

U.S. APPLICATION NO. (If known, see 37 CFR 1.53) <div style="font-size: 24pt; font-weight: bold; text-align: center;">10/049627</div>	INTERNATIONAL APPLICATION NO. PCT/SG99/00089																																																																																				
ATTORNEY'S DOCKET NUMBER P21834																																																																																					
19. <input checked="" type="checkbox"/> The following fees are submitted:																																																																																					
Basic National Fee (37 CFR 1.492(a)(1)-(5)): Search report has been prepared by the EPO or JPO. . . . . \$ 890.00 International preliminary examination fee paid to USPTO (37 CFR 1.482). . . . . \$ 710.00 No international preliminary examination fee paid to USPTO (37 CFR 1.482) but international search fee paid to USPTO (37 CFR 1.445(a)(2)). . . . . \$ 740.00 Neither international preliminary examination fee (37 CFR 1.482) nor international search fee (37 CFR 1.445(a)(2)) paid to USPTO. . . . . \$1,040.00 International preliminary examination fee paid to USPTO (37 CFR 1.482) and all claims satisfied provisions of PCT Article 33(2)-(4). . . . . \$ 100.00 <div style="text-align: center; margin-top: 10px;">ENTER APPROPRIATE BASIC FEE AMOUNT =</div>																																																																																					
Surecharge of \$130.00 for furnishing the oath or declaration later than ____ 20 ____ 30 months from the earliest claimed priority date (37 CFR 1.492(e)).																																																																																					
<table border="1" style="width:100%; border-collapse: collapse;"> <thead> <tr> <th style="width:15%;">Claims</th> <th style="width:15%;">Number Filed</th> <th style="width:15%;">Number Extra</th> <th style="width:15%;">RATE</th> <th style="width:15%;"></th> <th style="width:15%;"></th> </tr> </thead> <tbody> <tr> <td>Total Claims</td> <td>18 - 20 =</td> <td>0</td> <td>X \$18.00</td> <td>\$0.00</td> <td></td> </tr> <tr> <td>Independent Claims</td> <td>2 - 3 =</td> <td>0</td> <td>X \$84.00</td> <td>\$0.00</td> <td></td> </tr> <tr> <td colspan="3">Multiple dependent claim(s) (if applicable)</td> <td>+ \$280.00</td> <td>\$</td> <td></td> </tr> <tr> <td colspan="4" style="text-align: right;">TOTAL OF ABOVE CALCULATIONS =</td> <td>\$1040.00</td> <td></td> </tr> <tr> <td colspan="4"> <input type="checkbox"/> Applicant claims small entity status See 37 CFR 1.27. The fees indicated above are reduced            by 1/3.         </td> <td>\$</td> <td></td> </tr> <tr> <td colspan="4" style="text-align: right;">SUBTOTAL =</td> <td>\$1040.00</td> <td></td> </tr> <tr> <td colspan="4">           Processing fee of \$130.00 for furnishing the English translation later than ____ 20 ____ 30            months from the earliest claimed priority date (37 CFR 1.492(f)).         </td> <td>+</td> <td></td> </tr> <tr> <td colspan="4">Extension of Time fee in the amount of \$</td> <td></td> <td></td> </tr> <tr> <td colspan="4" style="text-align: right;">TOTAL NATIONAL FEE =</td> <td>\$1040.00</td> <td></td> </tr> <tr> <td colspan="4">           Fee for recording the enclosed assignment (37 CFR 1.21(h). The assignment must be            accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31). \$40.00 per property         </td> <td>+</td> <td></td> </tr> <tr> <td colspan="4" style="text-align: right;">TOTAL FEES ENCLOSED =</td> <td>\$1040.00</td> <td></td> </tr> <tr> <td colspan="4"></td> <td style="text-align: right;">Amount to be refunded</td> <td style="text-align: right;">\$</td> </tr> <tr> <td colspan="4"></td> <td style="text-align: right;">Charged</td> <td style="text-align: right;">\$</td> </tr> </tbody> </table>		Claims	Number Filed	Number Extra	RATE			Total Claims	18 - 20 =	0	X \$18.00	\$0.00		Independent Claims	2 - 3 =	0	X \$84.00	\$0.00		Multiple dependent claim(s) (if applicable)			+ \$280.00	\$		TOTAL OF ABOVE CALCULATIONS =				\$1040.00		<input type="checkbox"/> Applicant claims small entity status See 37 CFR 1.27. The fees indicated above are reduced by 1/3.				\$		SUBTOTAL =				\$1040.00		Processing fee of \$130.00 for furnishing the English translation later than ____ 20 ____ 30 months from the earliest claimed priority date (37 CFR 1.492(f)).				+		Extension of Time fee in the amount of \$						TOTAL NATIONAL FEE =				\$1040.00		Fee for recording the enclosed assignment (37 CFR 1.21(h). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31). \$40.00 per property				+		TOTAL FEES ENCLOSED =				\$1040.00						Amount to be refunded	\$					Charged	\$
Claims	Number Filed	Number Extra	RATE																																																																																		
Total Claims	18 - 20 =	0	X \$18.00	\$0.00																																																																																	
Independent Claims	2 - 3 =	0	X \$84.00	\$0.00																																																																																	
Multiple dependent claim(s) (if applicable)			+ \$280.00	\$																																																																																	
TOTAL OF ABOVE CALCULATIONS =				\$1040.00																																																																																	
<input type="checkbox"/> Applicant claims small entity status See 37 CFR 1.27. The fees indicated above are reduced by 1/3.				\$																																																																																	
SUBTOTAL =				\$1040.00																																																																																	
Processing fee of \$130.00 for furnishing the English translation later than ____ 20 ____ 30 months from the earliest claimed priority date (37 CFR 1.492(f)).				+																																																																																	
Extension of Time fee in the amount of \$																																																																																					
TOTAL NATIONAL FEE =				\$1040.00																																																																																	
Fee for recording the enclosed assignment (37 CFR 1.21(h). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31). \$40.00 per property				+																																																																																	
TOTAL FEES ENCLOSED =				\$1040.00																																																																																	
				Amount to be refunded	\$																																																																																
				Charged	\$																																																																																
a. <input checked="" type="checkbox"/> A check in the amount of \$1040.00 to cover the above fees is enclosed. b. <input type="checkbox"/> Please charge my Deposit Account No. _____ in the amount of \$ _____ to cover the above fees. c. <input checked="" type="checkbox"/> The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account No. 19-0089. NOTE: Where an appropriate time limit under 37 CFR 1.494 or 1.495 has not been met, a petition to revive (37 CFR 1.137(a) or (b)) must be filed and granted to restore the application to pending status.																																																																																					
SEND ALL CORRESPONDENCE TO CUSTOMER NO. 7055 AT THE PRESENT ADDRESS OF: Bruce H. Bernstein GREENBLUM & BERNSTEIN, P.L.C. 1941 Roland Clarke Place Reston, VA 20191 (703) 716-1191																																																																																					
 <div style="font-size: 24pt; font-weight: bold;">07055</div> PATENT TRADEMARK OFFICE																																																																																					
 SIGNATURE Bruce H. Bernstein NAME 29,027 REGISTRATION NUMBER																																																																																					

P21834.A01

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant : Ah Hwee TAN et al.

Serial No : Not Yet Assigned (National Stage of PCT/SG99/00089)

Filed : Concurrently Herewith (International Filing Date August 25, 1999)

For : DOCUMENT CLASSIFICATION APPARATUS

**PRELIMINARY AMENDMENT**

Commissioner of Patents and Trademarks  
Washington, D.C. 20231

Sir:

Prior to calculation of the filing fees and the examination of the above-identified patent application on the merits, the Examiner is respectfully requested to amend the claims as follows:

IN THE CLAIMS

Please amend the claims as follows (a marked-up copy of the claim amendments is provided as an attachment to this Amendment):

5. (Amended-Clean Text) Apparatus as claimed in claim 1 further comprising a router arranged to route the document to one of a plurality of destinations in dependence upon the classification.

P21834.A01

6. (Amended-Clean Text) Apparatus as claimed in claim 1 wherein the classification has associated therewith a confidence value.

7. (Amended-Clean Text) Apparatus as claimed in claim 1 further comprising a router arranged to route the document to one of a plurality of destinations in dependence upon the classification, wherein the classification has associated therewith a confidence value, and wherein the confidence value is comparable to a threshold, the router being arranged to make an automatic routing or manual routing decision in dependence upon the comparison.

9. (Amended-Clean Text) Apparatus as claimed in claim 7 or wherein a said destination is a system administrator, responsible for manual routing.

10. (Amended-Clean Text) Apparatus as claimed in claim 1 wherein the features are formed into a feature vector for input to the classifier.

11. (Amended-Clean Text) Apparatus as claimed in claim 1 wherein the features comprise classification-associated words or phrases which may appear in the document.

P21834.A01

12. (Amended-Clean Text) Apparatus as claimed in claim 1 wherein the extracting means is arranged to provide a measure of the frequency of occurrence of the features in the document.

15. (Amended-Clean Text) Apparatus as claimed in claim 1 wherein the apparatus is operable to perform rule insertion in the knowledge acquisition mode in which a plurality of features are input by a user to the classifier together with a classification with which the features are associated.

16. (Amended-Clean Text) Apparatus as claimed in claim 1 wherein the apparatus is operable in knowledge acquisition mode to process a plurality of training documents with associated classifications as a batch.

Please renumber the second claim 17 as claim 18 as follows:

18. (Amended-Clean Text) Apparatus as claimed in claim 13 wherein the threshold is adjustable.


P21834.A01

REMARKS

By the above amendment, claims 5, 6, 7, 9, 10, 11, 12, 15, and 16 have been amended to delete multiple dependency. The second claim 17 has been properly renumbered as claim 18.

If there should be any questions, the Examiner is invited to contact the undersigned at the telephone number listed below.

Respectfully submitted,  
Ah Hwee TAN et al.

  
Bruce H. Bernstein  
Reg. No. 29,027

*Reg. No.*  
33,329

February 21, 2002  
GREENBLUM & BERNSTEIN, P.L.C.  
1941 Roland Clarke Place  
Reston, VA 20191  
(703) 716-1191

P21834.A01

MARKED-UP COPY OF AMENDED CLAIMS

5. (Amended) Apparatus as claimed in claim 1 [any one of the preceding claims] further comprising a router arranged to route the document to one of a plurality of destinations in dependence upon the classification.

6. (Amended) Apparatus as claimed in claim 1 [any one of the preceding claims] wherein the classification has associated therewith a confidence value.

7. (Amended) Apparatus as claimed in claim 1 [6 as dependent on claim 5] further comprising a router arranged to route the document to one of a plurality of destinations in dependence upon the classification, wherein the classification has associated therewith a confidence value, and wherein the confidence value is comparable to a threshold, the router being arranged to make an automatic routing or manual routing decision in dependence upon the comparison.

9. (Amended) Apparatus as claimed in claim 7 or [claim 8] wherein a said destination is a system administrator, responsible for manual routing.

10. (Amended) Apparatus as claimed in claim 1 [any one of the preceding claims] wherein the features are formed into a feature vector for input to the classifier.

P21834.A01

11. (Amended) Apparatus as claimed in claim 1 [any one of the preceding claims] wherein the features comprise classification-associated words or phrases which may appear in the document.

12. (Amended) Apparatus as claimed in claim 1 [any one of the preceding claims] wherein the extracting means is arranged to provide a measure of the frequency of occurrence of the features in the document.

15. (Amended) Apparatus as claimed in claim 1 [any one of the preceding claims] wherein the apparatus is operable to perform rule insertion in the knowledge acquisition mode in which a plurality of features are input by a user to the classifier together with a classification with which the features are associated.

16. (Amended) Apparatus as claimed in claim 1 [any one of the preceding claims] wherein the apparatus is operable in knowledge acquisition mode to process a plurality of training documents with associated classifications as a batch.

18. [17] (Amended) Apparatus as claimed in claim 13 [16] wherein the threshold is adjustable.



General text classifying systems which classify documents into one or more categories have been proposed in US 5,371,807 and US 5,675,710. Such systems use only a single classification strategy, either profile-based, having a keyword/character profile for each category or rule-based in which category knowledge is represented in the form of rules. The systems also use only a single knowledge acquisition strategy, either statistically learned knowledge or user-specified knowledge to provide the knowledge base with which text from a document to be classified is compared to provide the document classification.

It is a disadvantage of the prior art systems noted above that they are prone to misclassification and consequent mis-routing of documents, as well as cumbersome operation.

- 5 It is an object of the invention to provide an improved document classification apparatus.

### SUMMARY OF THE INVENTION

According to the invention, there is provided document classification apparatus  
10 comprising feature extraction means for extracting a plurality of features from a document and a classifier operable on the extracted features to process the document in a knowledge acquisition mode in which the association of a classification with the document is added incrementally to a knowledge base or in a document classification mode in which the classifier, using the knowledge base, determines a predicted classification for the  
15 document, the classifier being switchable between the modes under user control .

The features are preferably formed into a feature vector for input to the classifier and the features preferably comprise classification-associated words or phrases which may appear in the document. The extracting means may be arranged to provide a measure of the  
20 frequency of occurrence of the features in the document.

The classifier may comprise a supervised ART system, preferably an ARAM system of the type disclosed in "Adaptive Resonance Associative Map", an article by one of the

present inventors Ah-Hwee Tan, published in "Neural Networks", Vol 8 No 3 pp 437-446. 1995 or an an ARTMAP system of the type disclosed in US 5,214,715.

The apparatus may further be operable in knowledge acquisition mode to process a plurality of training documents with associated classifications as a batch.

The apparatus may further be operable in a rule insertion sub-mode of the knowledge acquisition mode in which a plurality of features are input by a user to the classifier together with a classification with which the features are associated.

10

The apparatus may further comprising a router arranged to route the document to one of a plurality of destinations in dependence upon the classification and the classification may have associated therewith a confidence value comparable to a threshold, the router being arranged to make an automatic routing or manual routing decision in dependence upon the comparison, with a said destination being a system administrator, responsible for manual routing.

The described embodiment provides a document classification apparatus which allows learning to be performed in an incremental way by allowing a system administrator to correct document classification mistakes as they occur, the apparatus learning from these mistakes. By incremental learning of new cases does not require re-learning of previous cases, thus eliminating the need to preserve past cases for re-learning. While the described embodiment focuses primarily on incremental learning, the apparatus is further able to perform learning of a plurality of cases as a batch. During batch learning, the apparatus

learns each case one by one and accumulates the classification information into the knowledge base. Besides learning from training data, the apparatus also allows rules to be inserted into the learning process, leading to a more flexible learning environment.

- 5 The apparatus is furthermore able to determine a confidence that the classification of a particular document is correct in the form of a confidence value. This confidence value is compared to a threshold parameter to decide if automatic or manual routing is desirable. Adjustment of this threshold parameter allows the degrees of manual and automatic routing to be controlled, by adjustment of the threshold to match a desired confidence
- 10 value, thus allowing a smooth transition from a state where manual routing is favoured to one, as the classifier becomes more accurate, that favours automatic routing.

**BRIEF DESCRIPTION OF THE DRAWINGS**

Embodiments of the invention will now be described by way of example with reference to the accompanying drawings in which:

5 Figure 1 is a schematic diagram illustrating the structure of the described embodiments of the invention;

Figure 2 is a diagram illustrating the document classifier of Figure 1 in a document classification made;

10

Figure 3 is a diagram illustrating the modes of operation of the embodiments of the invention;

Figure 4 is a diagram of an ARAM system used as a document classifier in an  
15 embodiment of the invention.

Figures 5, 6, and 7 summarize the parameter setting and the relevant functional blocks of the document classification system in the learning, rule insertion, and document classification modes respectively.

20

**DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT**

With reference to Figures 1-3, a document classification apparatus is shown. The apparatus is operable in a knowledge acquisition mode and a document classification

mode. In knowledge acquisition mode, the apparatus learns from training documents and rules to recognise categories based on document content. This knowledge is then applied in document classification mode to classify further documents. The structure of the apparatus is shown in Figures 1 and 2 and will now be described with reference to the document classification mode. The structure is knowledge acquisition mode in the same, but used differently as described with reference to Figure 3.

A document text file, for example a text file derived from a scanned and OCR processed physical document or derived from a received and stored facsimile message which has been analysed for and converted to textual content, or a word processor document file, is fed to a document classifier 10. The document classifier includes a feature extraction module 20 which analyses the text file and extracts previously selected features in the form of keywords or phrases from that file which are fed as a feature vector to a classifier 30 which is in the form of an ARAM (Adaptive Resonance Associative Map) system which provides a predicted classification for the output document in response to the input feature vector. This classification is associated with a confidence value which, together with the document, is passed to a router 40. At the router 40, the value is compared to a threshold input by a system administrator 50. If the value exceeds the threshold, the document is routed to the destination 52 specified by the classification, via path 55. If not, the document is routed to the system administrator 50 for manual routing via path 60. The destinations 52 can also communicate with the system administrator 50 through path 60, to return misdirected documents for manual routing.

The modes of operation of the apparatus are shown in Figure 3. In **knowledge acquisition mode**, two sub-modes are used. The first, represented by block 100, is based on **learning** and requires the input of training data in the form of documents, for each of which a feature vector is extracted by module 20 and fed to module 30. The training documents can either be input individually or as a batch. The actual category of the document is input by system administrator 50 and fed to the module 30. Module 30 then adjusts (if necessary) the association of the vector to the predicted category in a knowledge base 105 so that the predicted category equals the actual category. The second sub-mode is based on **rule insertion**, represented by block 110. In rule insertion, a feature vector and an actual category are input by the system administrator 50 and an association between the input vector and the actual category is made, if one does not already exist.

In **document classification mode**, represented by block 120, the feature vector and document are fed to the module 30 and based on the knowledge acquired by the knowledge base in the knowledge acquisition mode, a classification is determined in accordance with the feature vector and the classification is output together with the document.

The system administrator can access the document classifier directly by via path 70 to allow switching between the knowledge acquisition sub-modes and the document classification mode. Such switching may be used, for example, if a mis-directed document has been returned to the system administrator. The system administrator may then cause the document classifier to enter the learning sub-mode of knowledge acquisition mode, the system administrator inputting the correct classification for the document to the

classifier 30 together with the document to the feature extraction module 20, from which the features are extracted and passed to the classifier 30, so that the mis-directed document and associated correct classification are added to the knowledge base.

- 5 Similarly, at any point in operation of the document classification apparatus, the systems administrator can add additional training documents and/or rules by switching from document classification mode to knowledge acquisition mode.

The highlighted processes will now be explained in more detail.

10

### Feature Selection

For document classification, there is a need to represent text documents in some format- and language- independent form, commonly termed a *feature* representation, before  
15 processing by a classifier. One of the most common forms of representation of features is that of singular word tokens. Specifically, the tokens are individual words that have been extracted from each document and transformed to their root form (e.g. root form of "selection" is "select"). Other "filtering" options based on sentence structure, such as recognizing only nouns while ignoring other word types such as prepositions and  
20 conjunctions, can also be used as will be apparent to those skilled in the art, in dependence upon requirements.

The keyword-based feature sets can be pre-defined manually or generated automatically from a set of pre-labeled documents.



The algorithm for automatic keyword selection accepts a list of pre-classified (i.e. training) text documents which are analyzed, processing one document at a time. Processing involves the extraction of all nouns (in root form) from each document and recording the number of occurrences of each of these prospective keywords within each category as well as within each document. Based on a certain set of selection rules, an overall rating of the "quality" of each word as a keyword is calculated and the list of keywords sorted by this value. The top N keywords with the highest rating are then selected as the "feature space" to be used for representing all documents (training or otherwise). The algorithm uses four different selection rules in ranking keywords which are combined to form a selection rating ( $f_{\text{rating}}$ ). These are:

- (a) Class Entropy
- (b) Document Entropy
- (c) Relative Document Count
- 15 (d) Document Inclusion Rate

a) Class Entropy ( $f_{\text{CE}}$ ): this measures the distribution of a keyword's occurrence across the different categories. The more "polarized" the keyword's occurrence is towards a particular category, the more significant will the keyword be. This is because a keyword which occurs almost solely within one category and not at all in the others is much more likely to have some non-trivial association with the that category, as compared with a keyword which has a more even distribution across the categories.

The formula used to calculate class entropy for C different categories is:

$$f_{CE} = 1 - \sum_{i=1}^{I \times C} \left( \left( \text{Count}(i) / \sum_{j=1}^{I \times C} \text{Count}(j) \right) \times \log(\text{Count}(i)) \right)$$

where:

$\text{Count}(x)$  = Total number of occurrences of keyword in category  $x$

5

b) Document Entropy ( $f_{DE}$ ): this measures the distribution of a keyword's occurrence across the different documents in a particular category. The criteria for a good keyword here is the opposite of that for Class Entropy. Here, the keyword which is much more evenly distributed across the documents in one category is a much better feature than one that has a more "polarized" distribution. This is because a keyword that occurs in more documents within a category is more likely to be one more commonly associated with documents of that category.

10

The formula used to calculate document entropy for D documents within 1 category, is:

$$f_{DE} = \sum_{i=1}^{I \times D} \left( \left( \text{Count}(i) / \sum_{j=1}^{I \times D} \text{Count}(j) \right) \times \log(\text{Count}(i)) \right)$$

15

where:

$\text{Count}(x)$  = Total number of occurrences of keyword in document  $x$

c) Relative Keyword Count ( $f_{RKC}$ ): for a particular keyword, the top 2 document categories are defined as the 2 categories with the highest absolute count for that keyword. The

20

keyword-per-document ratio ( $f_{\text{Ratio } i}$ ) for a category,  $i$ , is the total keyword count ( $C_i$ ) for the category divided by the total number of documents ( $D_i$ ) in that category. This relation can be expressed simply as:

$$f_{\text{Ratio } i} = C_i / D_i$$

Relative Keyword Count thus gives an indication of the difference between the keyword-per-document ratio of the 1<sup>st</sup> ( $f_{\text{Ratio } 1}$ ) and 2<sup>nd</sup> ( $f_{\text{Ratio } 2}$ ) categories. A keyword with a large difference between  $f_{\text{Ratio } 1}$  and  $f_{\text{Ratio } 2}$  is better than one with a small difference.

10

A measurement of  $f_{\text{RKC}}$  for  $C$  different categories is given by:

$$f_{\text{RKC}} = (f_{\text{Ratio } 1} - f_{\text{Ratio } 2}) / f_{\text{Ratio } 1}$$

d) Document Inclusion Rate ( $f_{\text{DIR}}$ ):  $f_{\text{RDC}}$  can be skewed by the high number of occurrences of a keyword in just one or two documents of a category. The use of  $f_{\text{DIR}}$  helps to "balance out" such situations by considering the number of documents in the top category in which the keyword occurs at least once.

15

A measurement of  $f_{\text{DIR}}$  for  $D_{\text{1st}}$  documents in the top category is given by:

$$f_{\text{DIR}} = O_{\text{1st}} / D_{\text{1st}}$$

20 where:

$O_{\text{1st}}$  = number of documents in top category in which keyword occurs.

The overall ranking of each keyword is therefore simply derived by taking:

$$f_{\text{Ranking}} = f_{\text{CE}} \times f_{\text{DE}} \times f_{\text{RDC}} \times f_{\text{DIR}}$$

with:

$$0.0 \leq f_{\text{Ranking}} \leq 1.0$$

In this case, equal weightage has been given to each factor. Different coefficients could easily be added to each factor to give it a larger or smaller weightage.

10 The following example uses a small training set of two categories with 124 relevant documents each. The categories are business newspaper articles in the first category and non-business (e.g. sports) articles in the other. Consider a sampling of 40 keywords taken from the set of all keywords selected from the training sets. The total count of each keyword within each category as well as the number of documents (per category) in which  
15 it occurs, is as shown in Table 1. In Table 2, the "paths" (as shown by arrowed lines) of two exemplary keywords as they are ranked according to the four different factors, together with the final rating are shown with the combination of the four factors helping to provide a better overall view of the relative suitability of each keyword.

**TABLE 1**

Key word	Total count		Unit document count	
	Cat 1	Cat 2	Cat 1	Cat 2
annual	33	2	22	2
authority	27	2	23	2
bank	92	6	28	4
capital	61	1	29	1
cent	560	32	87	21
champion	0	60	0	31
coach	0	26	0	19
company	191	12	65	8
corporate	48	4	26	3
cup	0	54	0	20
distribute	17	0	15	0
economy	87	1	28	1
event	2	34	1	25
exchange	34	2	23	2
fan	1	36	1	18
final	12	98	10	45
game	1	66	1	32
industry	137	6	59	6
invest	99	2	46	2
mainboard	26	0	22	0
market	175	4	64	3
match	6	65	3	28
minister	41	3	23	3
pc	499	43	91	21
play	12	181	11	55
potential	31	0	23	0
profit	78	5	28	3
property	96	3	37	3
rate	113	2	34	2
round	0	60	0	28
score	0	30	0	17
share	217	8	47	8
star	0	49	0	23
stock	104	0	40	0
technology	50	0	26	0
tournament	0	34	0	19
venture	53	1	20	1
victory	0	30	0	20
win	9	149	5	60
Woman	0	30	0	22

[illegible]

The algorithm allows for the specification of a minimum number  $K$ , of non-zero keyword counts which are expected to be found within each training document. The training documents are pre-processed by the method described above to determine the number of  
5 non-zero keyword counts in each document. Whenever a training document is found to have too few non-zero keyword counts, the next highest ranked keywords within the document are added to the set of  $N$  keywords initially selected, to bring the number of non-zero keyword counts for that document up to  $K$ . The total number of unique "bonus" keywords  $B$ , extracted from all training documents thus increases the dimension of the  
10 feature space to  $N+B$ .

### Keyword Extraction

Once the keywords have been selected in the manner described above, keywords are  
15 extracted from a document and are formed into a feature vector, using the  $N+B$  set of keywords obtained during the selection process as the limited set of significant keywords that are to be searched within the document. This procedure is applied to both training documents to produce a set of respective training feature vectors and new documents to produce, respective feature vectors for yet-to-be categorized documents.

20

Based on the selected keyword features, the feature extraction algorithm parses the document to record the number of times a keyword  $w_i$  appear in the document ( $c_i$ ). The keyword counts are then normalized such that the maximum score is 1 and the minimum score is 0. These scores are then provided as input to the classifier 30 as a normalized

keyword frequency count feature vector which encodes the statistical distribution of the keywords in the documents and thus provides a rough representation of the document content.

- 5 The feature extraction process using two sample articles is illustrated below. The first article, for Category 1 (business section) produces a positive word count for certain predominantly business-related keywords which are converted to relative frequency values as shown to form the input vector. The second article, for Category 2 (sports, music and life section) produces a positive word count for certain predominantly sports-related
- 10 keywords which are likewise converted to relative frequency values to form the input vector.



## Sample article for Category 1 (business section)

JUN 30 1997 Stationery maker Nippecraft in the red

MAINBOARD-LISTED specialist stationery maker Nippecraft has reported losses of \$12 million for the year ended March, but said a company reorganisation would improve its bottomline this year.

The losses came on the back of a 4 per cent drop in turnover to \$64 million and include exceptional and ordinary charges totalling close to \$11 million, according to the company's unaudited results.

There will be no dividend payouts this year. Net tangible asset backing per ordinary share dropped to 0.69 cent, from 7.82 cents last year. The results were in line with Nippecraft's projections announced in February.

Managing director Bill Habergham attributed a large part of the loss to reorganisation of businesses in Britain, the United States, Australia and Malaysia.

"The exercise has now largely been completed and notwithstanding the tougher prospects ahead, we expect to reap the benefits of the reorganisation and restructure exercise in the current financial year," a company statement quoted him as saying.

Nippecraft said the exceptional charges, amounting to \$6.8 million, included the writing down of stock by \$5 million, operating losses and costs associated with the closure or restructuring of subsidiaries which cost close to \$2 million.

The group managed to reduce inventory levels by a third to \$18 million. Mr Habergham said this would benefit the group in the long term.

Keyword Table for Sample Article for Category 1

Keyword	Count	Relative Frequency
market	0	0.0
cent	3	1.0
pe	2	0.7
industry	0	0.0
company	3	1.0
invest	0	0.0
develop	0	0.0
stock	1	0.3
share	1	0.3
list	1	0.3
property	0	0.0
technology	0	0.0
capital	0	0.0
economy	0	0.0
sector	0	0.0
billion	0	0.0
potential	0	0.0
mainboard	1	0.3
project	1	0.3
play	0	0.0
win	0	0.0
champion	0	0.0
game	0	0.0
rate	0	0.0
round	0	0.0
star	0	0.0
final	0	0.0
woman	0	0.0
victory	0	0.0
tournament	0	0.0

## Sample article for Category 2 (sports, music &amp; life section)

JUL 2 1997 Love-fit Testud clinches famous win

LONDON -- One point short of a famous victory, Sandrine Testud rolled her eyes to the leaden skies then across the net to Monica Seles, shifting nervously from foot to foot.

Finally she turned to her Italian boyfriend Vittorio, huddled among the spectators overlooking the No. 3 court.

On his signal she served straight, deep and fast for her sixth ace and an astonishing 0-6, 6-4, 8-6 win over the Wimbledon second seed.

Nothing had looked less likely half-an-hour into Monday's third-round match.

On a court of low and uncertain bounce after the heavy rain which ravaged the opening week, Seles breezed through the first set in exactly 30 minutes.

"I was just trying to start playing," said the unseeded Testud.

"I was just so slow and nothing was going right."

The 25-year-old Frenchwoman, who lives and trains in Rome, finally won a game, holding serve in the second game of the second set after dropping the first two points.

She broke Seles in the next game. Hitting longer and with more power as she gained in confidence, Testud held service to win the second set in 42 minutes.

"I got a little bit tight, missed a couple of shots and the set was gone," Seles reflected.

Vittorio's contribution apparently goes further than his court-side advice.

How have you gotten so fit?, Testud was asked.

Keyword Table for Sample Article for Category 2

Keyword	Count	Relative Frequency
Market	0	0.0
Cent	0	0.0
Pe	0	0.0
Industry	0	0.0
Company	0	0.0
Invest	0	0.0
develop	0	0.0
stock	0	0.0
share	0	0.0
list	0	0.0
property	0	0.0
technology	0	0.0
capital	0	0.0
economy	0	0.0
sector	0	0.0
billion	0	0.0
Potential	0	0.0
Mainboard	0	0.0
Project	0	0.0
Play	0	0.0
Win	4	1.0
Champion	0	0.0
Game	3	0.8
rate	0	0.0
round	1	0.2
star	0	0.0
final	2	0.5
woman	0	0.0
victory	1	0.2
tournament	0	0.0

The Classifier: Adaptive Resonance Associative Map (ARAM)

ARAM is a family of neural network models that performs incremental supervised learning of recognition categories (pattern classes) and multidimensional maps of both binary and analog patterns. An ARAM system is shown in Figure 4 and can be visualized as two overlapping Adaptive Resonance Theory (ART) [1,2,3] modules consisting of two input fields  $F_1^a$  (300) and  $F_1^b$  (310) with an  $F_2$  category field (320). For classification problems, the  $F_1^a$  field (300) serves as the input field containing the input activity vector and the  $F_1^b$  field (310) servers as the output field containing the output class vector. The  $F_2$  field (320) contains the activities of categories that are used to encode the patterns. During learning, given an input pattern presented at the  $F_1^a$  input layer and an output pattern presented at the  $F_1^b$  output field, a  $F_2$  category node is selected to encode the pattern pair.

When performing classification tasks, ARAM formulates recognition categories of input patterns, and associates each category with its respective prediction. The knowledge that ARAM discovers during learning is compatible with IF-THEN rule-based representation. Specifically, each node in the  $F_2$  field (320) represents a recognition category associating the  $F_1^a$  input patterns with the  $F_1^b$  output vectors. Learned weight vectors, one for each  $F_2$  node, constitute a set of rules that link antecedents to consequents. At any point during the incremental learning process, the system architecture can be translated into a compact set of rules. Similarly, domain knowledge in the form of IF-THEN rules can be inserted into ARAM architecture.

The ART modules used in ARAM can be ART 1 [1], which categorizes binary patterns, or analog ART modules such as ART 2-A [2], and fuzzy ART [3], which categorize both binary and analog patterns. The fuzzy ARAM model, that is composed of two overlapping fuzzy ART modules is described below.

5

### Knowledge Acquisition Mode

#### Learning Sub-Mode

In the learning sub-mode of knowledge acquisition mode, ARAM learns a set of  
10 recognition categories or rules by training from pre-labeled document sets. During learning, the keyword frequency vectors, each representing a document, are presented to ARAM as input A one at a time together with the associated class label input B.

Given an input keyword vector A, ARAM first searches for a  $F_2$  recognition category  
15 encoding a keyword template vector that is closest to the input vector according to some similarity measure. It then checks if the associated  $F_2$  output prediction of the selected category matches with the output class label B. If so, under fast learning, the keyword templates of the  $F_2$  recognition category is modified to contain the intersection of the original keyword templates and the input keyword vector. Otherwise, the recognition  
20 category is reset and the system repeats to select another category until a match is found.

Given a set of the documents with a specific class label, the system learns to pick up combinations of keywords that consistently appear in the documents and derive rules that associate combinations of the relevant keywords to the target output class of the

documents. ARAM learning is stable in the sense that weight values do not oscillate, as they can only decrease but not increase. As new cases are incorporated by adjusting the weight templates of the chosen category nodes, learning does not wash away previously learned knowledge. This allows incremental learning in the sense that learning of new cases does not require relearning of old data.

The detailed algorithm for incremental learning is given below:

Input vectors: The  $F_1^a$  and  $F_1^b$  input vectors are normalized by *complement coding* that preserves amplitude information. Complement coding represents both the on-response and the off-response to an input vector  $a$ . The complement coded  $F_1^a$  input vector  $A$  is a  $2M$ -dimensional vector

$$A = (a, a^c) = (a_1, \dots, a_M, a_1^c, \dots, a_M^c)$$

Where  $a_i$  represents the normalized frequency score of keyword  $w_i$ , and  $a_i^c = 1 - a_i$ .

Similarly, the complement coded  $F_1^b$  input vector  $B$  is a  $2N$ -dimensional vector

$$B = (b, b^c) = (b_1, \dots, b_N, b_1^c, \dots, b_N^c)$$

where  $b_k$  represents the presence ( $b_k=1$ ) or absence ( $b_k=0$ ) of a class label  $c_k$ , and  $b_k^c = 1 - b_k$ .

Activity vectors: Let  $x^a$  and  $x^b$  denote the  $F_1^a$  and  $F_1^b$  activity vectors respectively. Let  $y$  denote the  $F_2$  activity vector.

Weight vectors: Each  $F_2$  category node  $j$  is associated with two adaptive weight templates  $w_j^a$  and  $w_j^b$ . Initially, all category nodes are uncommitted and all weights equal ones. After a category node is selected for encoding, it becomes committed.

Fuzzy ARAM dynamics are determined by the choice parameters  $\alpha_a > 0$  and  $\alpha_b > 0$ ; the learning rates  $\beta_a$  in  $[0,1]$  and  $\beta_b$  in  $[0,1]$ ; the vigilance parameters  $\rho_a$  in  $[0,1]$  and  $\rho_b$  in  $[0,1]$ ; and a contribution parameter  $\gamma$  in  $[0,1]$ .

Bottom up propagation: Given the  $F_1^a$  input vectors  $A$ , for each  $F_2$  node  $j$ , the  $F_1^a$  to  $F_2$  input  $T_j$  is defined by :

$$T_j = |A \wedge w_j^a| / (\alpha_a + |w_j^a|)$$

where the fuzzy AND operation  $\wedge$  is defined by  $(p \wedge q)_i = \min(p_i, q_i)$ , and where the norm  $|\cdot|$  is defined by  $|p| = \sum_i p_i$  for vectors  $p$  and  $q$ .

Category choice: Using a choice rule, at most one  $F_2$  node can become active. The choice is indexed at  $J$  where  $T_J = \max \{T_j; \text{ for all } F_2 \text{ node } j\}$ .

When a category choice is made at node  $J$ ,  $y_J = 1$ ; and  $y_j = 0$  for all  $j$  not equal to  $J$ .



Resonance or reset: Resonance occurs if the *match functions*,  $m_j^a$  and  $m_j^b$ , meet the vigilance criteria in their respective modules:

$$5 \quad m_j^a = |A \wedge w_j^a| / |A| \geq \rho_a \text{ and } m_j^b = |B \wedge w_j^b| / |B| \geq \rho_b.$$

Learning then ensues, as defined below. If any of the vigilance constraints is violated, mismatch reset occurs in which the value of the choice function  $T_j$  is set to 0 for the duration of the input presentation. The search process repeats to select another new index

10  $J$  until resonance is achieved.

Learning : Once the search ends, the weight vectors  $w_j^a$  and  $w_j^b$  are updated according to the equations

$$15 \quad w_j^{a \text{ (new)}} = (1-\beta_a) w_j^{a \text{ (old)}} + \beta_a (A \wedge w_j^{a \text{ (old)}})$$

and

$$w_j^{b \text{ (new)}} = (1-\beta_b) w_j^{b \text{ (old)}} + \beta_b (B \wedge w_j^{b \text{ (old)}})$$

respectively. For efficient coding of noisy input sets, it is useful to set  $\beta_a = \beta_b = 1$  when

20  $J$  is an uncommitted node, and then take  $\beta_a < 1$  and  $\beta_b < 1$  after the category node is committed. *Fast learning* corresponds to setting  $\beta_a = \beta_b = 1$  for committed nodes.

Match tracking: At the start of each input presentation, the vigilance parameter  $\rho_s$  equals a baseline vigilance  $\rho_s$ . If a reset occurs in the category field  $F_2$ ,  $\rho_s$  is increased until it is slightly larger than the match function  $m_{ij}$ . The search process then selects another  $F_2$  node  $J$  under the revised vigilance criterion.

5

### Rule Insertion Sub-Mode

Through the rule insertion process, user-defined knowledge in the form of rules is inserted into the ARAM network (knowledge base). A rule is typically in the IF-THEN format, consisting of a set of keyword features as the antecedents and a classification as the consequent. Due to the compatibility of ARAM architecture and rules, domain knowledge in the form of IF-THEN rules can be readily inserted into an ARAM network.

For example, given a rule such as

15

*"Stock", "Share", "Price" -> Business,*

the rule insertion algorithm creates a keyword frequent vector in which the frequency score of *"Stock"*, *"Share"* and *"price"* are 1s and all others zeros; and an output class vector in which the score of *"Business"* is 1 and all others zeros. Given the keyword frequency vector in the  $F_1^a$  field, and the class vector in the  $F_1^b$  field, ARAM first searches for a recognition category that encodes the exact same set of keywords. If such a recognition category exists and its predicted class is *"Business"*, no action is required as the rule already exists. If the predicted class is not *"Business"*, a contradiction occurs and it is flagged to the users. If such a recognition category does not exist, a recognition

20

category is created to encode a keyword template consisting of "Stock", "Share", and "Price" and a classification of "Business".

The detailed rule insertion algorithm is as follows:

5

Rule insertion proceeds in two phases. The first phase translates each rule into a  $2M$ -dimensional vectors **A** and a  $2N$ -dimensional vectors **B**, where  $M$  is the total number of document features and  $N$  is the number of classes.

10 In the most general case, given a rule of the following format,

IF  $x_1, x_2, \dots, x_m, \text{not}(x'_1), \text{not}(x'_2), \dots, \text{not}(x'_m)$

THEN  $y_1, y_2, \dots, y_n, \text{not}(y'_1), \text{not}(y'_2), \dots, \text{not}(y'_m)$

15

where  $x_1, x_2, \dots, x_m$  and  $y_1, y_2, \dots, y_n$  are positive attributes, and  $x'_1, x'_2, \dots, x'_m$  and  $y'_1, y'_2, \dots, y'_n$  preceded by the logical NOT operator are negative attributes, the algorithm derives the pair of vectors

20

$$\mathbf{A} = (\mathbf{a}, \mathbf{a}^c) \text{ and } \mathbf{B} = (\mathbf{b}, \mathbf{b}^c)$$

such that for each index  $j = 1, \dots, M$ ,

$$(\mathbf{a}_j, \mathbf{a}_j^c) =$$

(1,0) if  $w_j = x_i$  for some  $i$  in  $\{1, \dots, m\}$

(0,1) if  $w_j = x'_i$  for some  $i$  in  $\{1, \dots, m'\}$

(0,0) otherwise

and

5

$(b_k, b'_k) =$

(1,0) if  $c_k = y_i$  for some  $i$  in  $\{1, \dots, n\}$

(0,1) if  $c_k = y'_i$  for some  $i$  in  $\{1, \dots, n'\}$

(0,0) otherwise

10

where  $w_j$  is the  $j^{\text{th}}$  keyword feature and  $c_k$  is the  $k^{\text{th}}$  class label.

The vector pairs derived from the rules are then used as training patterns to initialize an ARAM network. Given a pair of vectors A and B derived from a rule, their respective

15 recognition categories are associated through the map field.

During network initialization, the vigilance parameters  $\rho_a$  and  $\rho_b$  are each set to 1 to ensure that only identical attribute vectors are grouped into one recognition category. Contradictory symbolic rules are detected during rule insertion when identical input attribute vectors are

20 associated with distinct output attribute vectors. The detection is achieved through a perfect mismatch phenomenon, in which the system tries to raise  $\rho_a$  above 1 in response to a mismatch in  $F_1^a$ .

Document classification

Given an input document, a feature extraction module parses the text to derive a normalized keyword frequency vector (as described above). The complement coded input  
 5 vector  $A$  is then presented to the  $F_1^a$  field.

Given an input keyword vector  $A$ , ARAM first searches for a  $F_2$  recognition category encoding a keyword template vector that is closest to the input vector according to the choice function. The associated  $F_2$  output prediction of the selected category is then used  
 10 as the output class label.

Choice Rule: In ARAM systems with category choice, only the  $F_2$  node  $J$  that receives maximal  $F_1^a$  to  $F_2$  input  $T_j$  predicts output. Specifically:

15  $y_j = 1$  if  $j = J$  where  $T_j > T_k$  for all  $k$  not equal to  $J$ ;

0 otherwise

The  $F_1^b$  activity vector  $x^b$  is given by  $x^b = \sum_j w_j^b y_j = w_j^b$  and the output vector  $B_2 = (b_1, b_2, \dots, b_n)$  is then read directly from the  $F_1^b$  field. The output class is predicted to be  $K$  if  $b_K > b_k$  for  
 20 all  $k$  not equal to  $K$  and the confidence value is given by  $b_K$ .

**Confidence Value**

Given training examples and rules of a single class output and with fast learning, ARAM  
 5 associates input features to a binary class prediction. In other words, only one output class  
 $b_k$  equals one and  $b_k = 0$  for all  $k$  not equal to  $K$ . To derive a real value prediction score  
 between 0 and 1, a few strategies are possible, of which two are described below:

**a) Distributed category prediction**

Using distributed category prediction, more than one  $F_2$  nodes can become active. The  $F_2$   
 10 output vector  $y$  represents a less extreme contrast enhancement of the  $F_1^2$  to  $F_2$  input  $T$ , in  
 the sense that the higher  $T_j$ 's are amplified and smaller  $T_j$ 's are further reduced. Two  
 algorithms that approximate contrast enhancement are given below.

**Power Rule:** The power rule raises the input  $T_j$  to the  $j^{\text{th}}$   $F_2$  node to a power  $p$  and  
 15 normalizes the total activity:

$$y_j = (T_j)^p / \sum_k (T_k)^p.$$

The power rule converges toward the choice rule as  $p$  becomes large.

20

**K-max Rule:** In the spirit of the K Nearest Neighbor (KNN) system, the K-max rule picks  
 the set of  $K$   $F_2$  nodes with the largest input  $T_j$  for prediction. The  $F_2$  activity values  $y_j$  are  
 then:

$$y_j = T_j / \sum_{k \in \pi} T_k \quad \text{if } j \in \pi$$

$$0 \quad \text{otherwise,}$$

- 5 where  $\pi$  is the set of  $K$  category nodes with the largest  $T_j$  values. The  $K$ -max rule with  $K=N$  is equivalent to the power rule with  $p=1$ .

After the  $F_2$  activity vector  $y$  is contrast enhanced by the power rule or the  $K$ -max rule, the output activity vector  $x^b$  in the  $F_1^b$  field computed by

10

$$x^b = \sum_j w_j^b y_j$$

The output vector  $B_2 = (b_1, b_2, \dots, b_N)$  is then read directly from  $x^b$ . The output class is predicted to be  $K$  if  $b_K > b_k$  for all  $k$  not equal to  $K$  and the confidence value is given by  $b_K$ .

15

#### b) Voting strategy

- Using the voting strategy technique, multiple ARAM systems are inserted with different sets of rules and/or trained using different sets of input patterns or different orderings of
- 20 the input patterns. When performing classification, each ARAM votes for its predicted class. The voting scores normalized by the number of ARAM provide a prediction score between 0 and 1 for each output class.

$$s_j = v_j / \sum_k v_k$$

where  $v_j$  is the number of votes given to and  $s_j$  is the normalized prediction score for the output class  $j$ . The output class with the highest prediction score is the selected predicted class and its prediction score is the confidence value.

### Switching between modes

The system administrator can switch between the classification mode and the knowledge acquisition sub-modes by sending a message together with the appropriate data to the document classifier. The message can be either LEARN, INSERT, or CLASSIFY. Depending on the message received, the document classifier adjusts the input baseline vigilance parameter  $\rho_a$  and the output vigilance parameter  $\rho_b$  of the ARAM classifier accordingly and carries out the appropriate sequence of actions.

15

With a LEARN message, the document classifier receives a document text together with a classification label. First, the feature extraction module derives the normalized keyword frequency vector from the document. The keyword vector is presented as the input vector to the  $F_1^a$  field and the classification vector (based on the classification label) is presented to the  $F_1^b$  field of the ARAM classifier. The ARAM classifier is then run with  $\rho_a < 1$  (typically 0, to maximize generalization) and  $\rho_b = 1$ .

20

With an INSERT message, the document classifier receives an IF-THEN rule. First, the rule insertion module converts the given rule into a pair of input and output vectors, presents the



input vector to the  $F_1^a$  field and the output vector to the  $F_1^b$  field. The ARAM classifier is then run with both the input and output vigilance parameters set to 1s.

With a CLASSIFY message, the document classifier receives a document text. First, the  
5 feature extraction module derives the normalized keyword frequency vector from the document and presents it as the input vector to the  $F_1^a$  field. The ARAM classifier is then run with both  $\rho_a$  and  $\rho_b$  equal to zeroes to ensure a prediction is always made. The predicted classification label is then read from the  $F_1^b$  field and returned to the user.

10 Figures 5, 6, and 7 summarize the parameter setting and the relevant functional blocks of the document classification system in the learning, rule insertion, and document classification modes respectively.

The embodiment described is not to be construed as limitative. For example, although the  
15 classifier module has been shown implemented using an ARAM structure, this may be implemented using any other structure which allows incremental learning and rule insertion, such as an ARTMAP structure.

## References:

[1] G. A. Carpenter & S. Grossberg (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37, 54-115.

[2] G. A. Carpenter, S. Grossberg & D. B. Rosen (1991a). ART 2-A: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4, 493-504.

[3] G. A. Carpenter, S. Grossberg & D. B. Rosen (1991b). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4, 759-771.

[4] C. A. Carpenter & A.-H. Tan (1995). Rule Extraction: From Neural Architecture to Symbolic Representation'. *Connection Science*, 7(1), 3-27.

[5] A.-H. Tan (1995). Adaptive Resonance Associative Map. *Neural Networks*, 8(3), 437-446.

[6] A.-H. Tan, "Cascade ARTMAP: Integrating Neural Computation and Symbolic Knowledge Processing". *IEEE Transactions on Neural Networks*, Vol. 8, No. 2 (March 1997) 237-250.

## CLAIMS

1. Document classification apparatus comprising feature extraction means for  
extracting a plurality of features from a document and a classifier operable on the  
5 extracted features to process the document in a knowledge acquisition mode in  
which the association of a classification with the document is added incrementally  
to a knowledge base or in a document classification mode in which the classifier,  
using the knowledge base, determines a predicted classification for the document,  
the classifier being switchable between the modes under user control .

10  
2. Apparatus as claimed in claim 1 wherein the classifier comprises a supervised  
adaptive resonance theory (ART) system.

3. Apparatus as claimed in claim 2 wherein the system comprises an ARTMAP  
15 system.

4. Apparatus as claimed in claim 2 wherein the system comprises an ARAM system.

5. Apparatus as claimed in any one of the preceding claims further comprising a  
20 router arranged to route the document to one of a plurality of destinations in  
dependence upon the classification.

6. Apparatus as claimed in any one of the preceding claims wherein the classification  
has associated therewith a confidence value.

7. Apparatus as claimed in claim 6 as dependent on claim 5 wherein the confidence value is comparable to a threshold, the router being arranged to make an automatic routing or manual routing decision in dependence upon the comparison.

5

8. Apparatus as claimed in claim 7 wherein the threshold is adjustable.

9. Apparatus as claimed in claim 7 or claim 8 wherein a said destination is a system administrator, responsible for manual routing.

10

10. Apparatus as claimed in any one of the preceding claims wherein the features are formed into a feature vector for input to the classifier.

15

11. Apparatus as claimed in any one of the preceding claims wherein the features comprise classification-associated words or phrases which may appear in the document.

20

12. Apparatus as claimed in any one of the preceding claims wherein the extracting means is arranged to provide a measure of the frequency of occurrence of the features in the document.

13. Apparatus as claimed in claim 5 wherein the destinations include a system administrator to which the other destinations are connected, mis-routed documents

being sendable by the other destinations to the system administrator for manual routing.

14. Apparatus as claimed in claim 13 wherein the system administrator is connected to  
5 the feature extraction means and classifier, the arrangement being such that a said mis-directed document, in association with an actual classification supplied by the system administrator, is processed in knowledge acquisition mode to add the association of the actual classification with the mis-directed document to the knowledge base.

10  
15. Apparatus as claimed in any one of the preceding claims wherein the apparatus is operable to perform rule insertion in the knowledge acquisition mode in which a plurality of features are input by a user to the classifier together with a classification with which the features are associated.

15  
16. Apparatus as claimed in any one of the preceding claims wherein the apparatus is operable in knowledge acquisition mode to process a plurality of training documents with associated classifications as a batch.

20 17. Document classification apparatus comprising:  
feature extraction means for extracting a plurality of features from a document,  
a classifier operable, using a knowledge base, to determine from the features a predicted classification for the document, the classification having a confidence value associated therewith; and

a router arranged to compare the confidence value to a threshold and make a decision to route the document automatically to one of a plurality of destinations or to a destination for manual routing in dependence upon the comparison.

- 5 17. Apparatus as claimed in claim 13 wherein the threshold is adjustable.



1/6

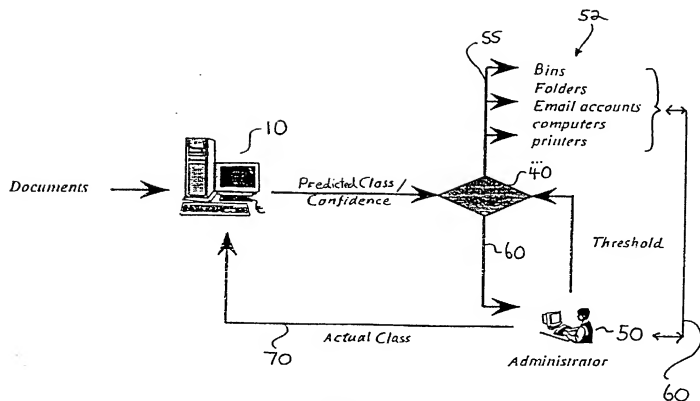


FIG. 1

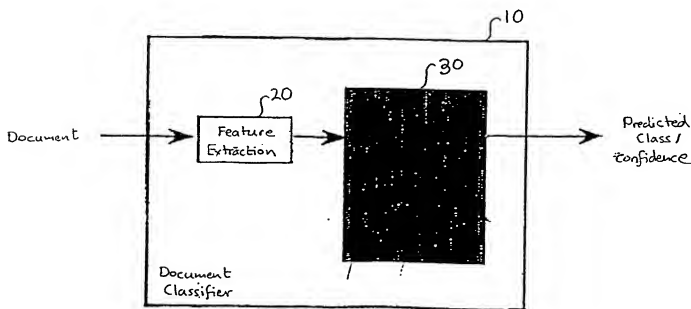


FIG. 2



2/6

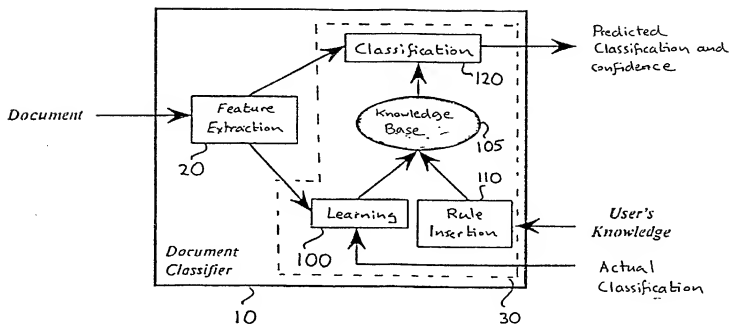


FIG. 3

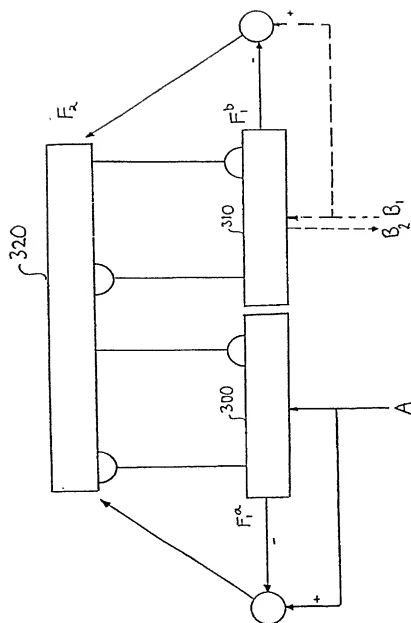


FIG. 4

ARAM Learning:  $p^a < 1, p^b = 1$

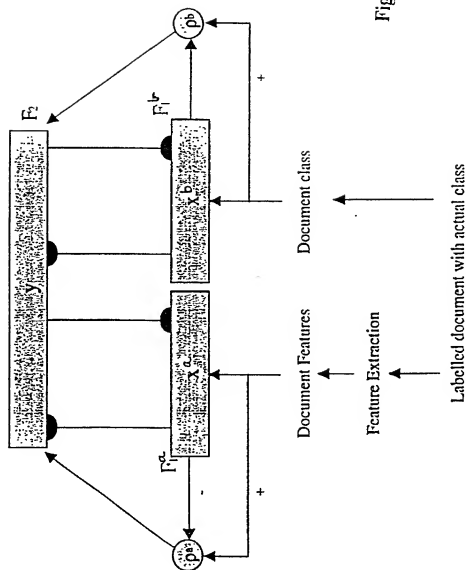


Fig. 5

5/6

ARAM rule insertion:  $p^a = 1, p^b = 1$

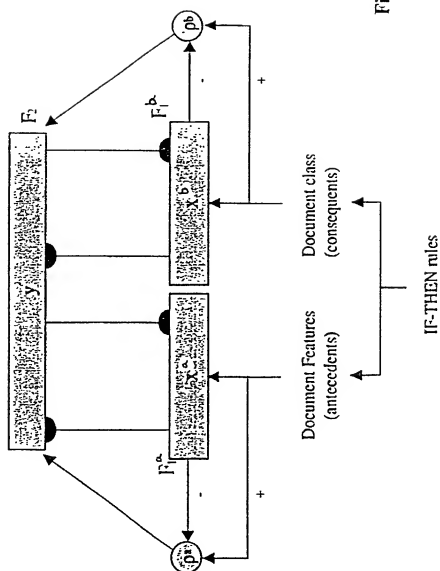


Fig. 6

ARAM classification:  $\rho^a = 0, \rho^b = 0$

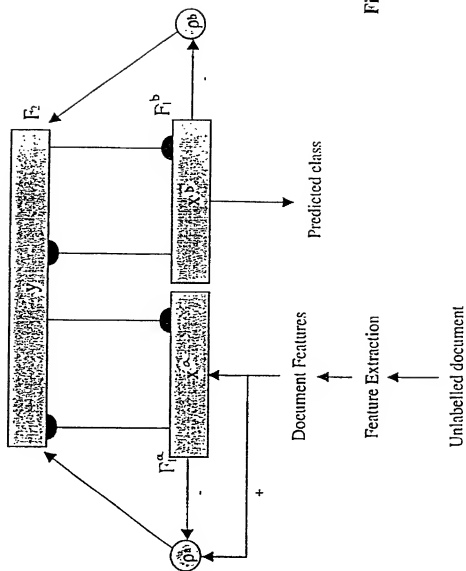


Fig. 7

# Declaration and Power of Attorney For Utility or Design Patent Application

## English Language Declaration

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled

### DOCUMENT CLASSIFICATION APPARATUS

the specification of which is attached hereto unless the following box is checked:

☒ was filed on 25 AUGUST 1999 as

United States Application Number \_\_\_\_\_ (if applicable) or,  
and was amended on \_\_\_\_\_

PCT International Application Number PCT/SG99/00089 (if applicable)  
and was amended on \_\_\_\_\_

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations, §1.56.

I hereby claim foreign priority benefits under Title 35, United States Code §119 (a-d) or §365(b) of any foreign application(s) for patent or inventor's certificate, or §365(a) of any PCT international application which designated at least one country other than the United States of America, listed below. I have also identified below, by checking the "No" box, any foreign application for patent or inventor's certificate, or of any PCT international application having a filing date before that of the application on which priority is claimed:

			Priority Claimed	
			<input type="checkbox"/>	<input type="checkbox"/>
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	Yes <input type="checkbox"/>	No <input type="checkbox"/>
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	Yes <input type="checkbox"/>	No <input type="checkbox"/>
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	Yes <input type="checkbox"/>	No <input type="checkbox"/>

☐ Additional foreign application numbers are listed on a supplemental priority sheet attached hereto.

I hereby claim the benefit under Title 35, United States Code §119(e) of any United States provisional application(s) listed below.

_____ (Number)	_____ (Day/Month/Year Filed)
_____ (Number)	_____ (Day/Month/Year Filed)
_____ (Number)	_____ (Day/Month/Year Filed)

☐ Additional provisional application numbers are listed on a supplemental priority sheet attached hereto.

I hereby claim the benefit under Title 35, United States Code §120 of any United States application(s), or §365(c) of any PCT international application designating the United States of America, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT international application in the manner provided by the first paragraph of Title 35, United States Code §112, I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations §1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application.

(Application No.)

(Filing Date)

(Status)  
(patented, pending, abandoned)

(Application No.)

(Filing Date)

(Status)  
(patented, pending, abandoned)

☐ Additional U.S. or international application numbers are listed on a supplemental priority sheet attached hereto.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

The undersigned hereby authorizes the U.S. attorney or agent named herein to accept and follow instructions from either his foreign patent agent or corporate representative, if any, as to any action to be taken in the Patent and Trademark Office regarding this application without direct communication between the U.S. attorney or agent and the undersigned. In the event of a change in the persons from whom instructions may be taken, the U.S. attorney or agent named herein will be so notified by the undersigned.

**POWER OF ATTORNEY:** As a named inventor, I hereby appoint the attorney(s) and/or agent(s) associated with the Customer Number provided below to prosecute this application and transact all business in the Patent and Trademark Office connected therewith, and direct that all correspondence be addressed to that Customer Number:

**CUSTOMER NUMBER 7055**

The appointed attorneys include:

Neil F. Greenblum Reg. No. 28,394  
Bruce H. Bernstein Reg. No. 29,027  
Arnold Turk Reg. No. 33,094  
James L. Rowland Reg. No. 32,674

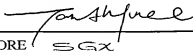
Stephen M. Roylance Reg. No. 31,296  
Leslie J. Paperner Reg. No. 33,329  
William Pieprz Reg. No. 33,630  
William E. Lyddane Reg. No. 41,568

At: Greenblum & Bernstein, P.L.C.  
1941 Roland Clarke Place  
Reston, VA 20191

Direct Telephone Calls to: Greenblum & Bernstein, P.L.C. (703) 716-1191

**Full name of sole or first inventor AH HWE E TAN**

Inventor's signature



Date 7 Feb 2002

Residence SINGAPORE

SGX

Citizenship SINGAPORE

Post Office Address 63 TAMPINES AVENUE 1, #05-01, THE TROPICA, SINGAPORE 529777

(Supply similar information and signature for second and subsequent joint inventors.)

2-05 <b>Full name of second joint inventor, if any</b> <u>FON LIN LAI</u>	
Second Inventor's signature <i>Carl Forster</i>	Date <u>7/2/2002</u>
Residence <u>SINGAPORE</u> <u>SSX</u>	
Citizenship <u>MALAYSIA</u>	
Post Office Address <u>101-G JERVOIS ROAD, SINGAPORE 249058</u>	
<b>Full name of third joint inventor, if any</b>	
Third Inventor's signature	Date
Residence	
Citizenship	
Post Office Address	
<b>Full name of fourth joint inventor, if any</b>	
Fourth Inventor's signature	Date
Residence	
Citizenship	
Post Office Address	
<b>Full name of fifth joint inventor, if any</b>	
Fifth Inventor's signature	Date
Residence	
Citizenship	
Post Office Address	